DOCUMENT RESUME

ED 334 202 TM 016 592

AUTHOR Mehrens, William A.
TITLE Defensible/Indefensible Instructional Preparation for High Stakes Achievement Tests: An Exploratory Trialogue.
PUB DATE 10 Apr 91
NOTE 12p.; Revision of a paper presented at the Annual Meetings of the American Educational Research Association (Chicago, IL, April 3-7, 1991) and the National Council on Measurement in Education (Chicago, IL, April 4-6, 1991).
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Elementary Secondary Education; Guidelines; *Instructional Effectiveness; *Standardized Tests; *Teacher Role; *Test Coaching; Test Use
IDENTIFIERS *High Stakes Tests; Teaching to the Test

ABSTRACT

Issues involved in high stakes testing are reviewed, with emphasis on the proper role of instructional preparation. The recent focus on educational accountability has increased pressure to raise test scores. One way of improving test scores is to teach what is on the test. The following guidelines concerning appropriate instructional strategies are presented: (1) a teacher should not engage in instruction that attenuates the ability to infer from the test score to the domain of knowledge/skill/ability of interest; (2) it is appropriate to teach the content domain to which the user wishes to infer; (3) it is appropriate to teach test-taking skills; (4) it is inappropriate to limit content instruction to a particular test item format; (5) it is inappropriate to teach only objectives from the domain that are sampled on the test; (6) it is inappropriate to use an instructional guide that reviews the questions of the latest issue of the test; (7) it is inappropriate to limit instruction to the actual test questions; (8) it is appropriate to teach toward test objectives if the test objective comprise the domain objectives; (9) it is appropriate to ensure that students understand the test vocabulary; and (10) one cannot teach only the specific task of a performance assessment. Grey areas and tangential issues in test preparation are discussed. (SLD)

Revised 4/10/91

DEFENSIBLE/INDEFENSIBLE INSTRUCTIONAL PREPARATION FOR HIGH STAKES ACHIEVEMENT TESTS: AN EXPLORATORY TRIALOGUE

William A. Mehrens
462 Erickson Hall
Michigan State University
East Lansing, MI 48824

(517) 355-9567

Revision of a presentation given in a symposium (same title) with N. S. Cole and W. J. Popham at the 1991 AERA/NCME Annual Meetings, Chicago.

DEFENSIBLE/INDEFENSIBLE INSTRUCTIONAL PREPARATION FOR HIGH STAKES ACHIEVEMENT TESTS: AN EXPLORATORY TRIALOGUE

William A. Mehrens

The purpose of the symposium is to generate, then judge, the suitability of various test-preparation options. My original task is to set the stage for an analytic discussion rather than to provide definitive answers to the question posed in the title of this symposium. To set the stage, I will (1) give you my version of the historical bases of the issue, (2) review some psychometricians' writings on the issue, (3) present my current views on defensible/indefensible instructional preparation for high stakes achievement tests, (4) mention some grey areas, and (5) discuss the following set of somewhat tangential issues: (a) whether we should have high-stakes tests, (b) what are the responsibilities of those who mandate and/or build the tests, and (c) whether this issue is important for performance assessment.

HISTORICAL BASES OF THE ISSUE

My remarks on the historical bases of the issue can be divided into two sub-headings: (1) Why educators are confused and (2) why the public is concerned.

Educators are confused because for decades they had been told that one should not teach the test. More recently, however, critics of tests "discovered" that all tests do not measure the same thing and that one should obtain curricular/test alignment either by changing the test and/or the curriculum. Further, in recent years the public has wished to hold educators accountable for student learning and has felt that test scores were the best data one could obtain regarding student levels of achievement. This accountability movement has increased pressure to raise test scores.

Obviously, one way to do that is to teach what is on the test. The logic seems acceptable to many. It goes like this: If I (the teacher) am to be held accountable for my students learning a specific set of things, it is certainly in my best interest to teach those specific things. The problem, of course, is whether by a specific set of things, one is discussing the particular sample of objectives tested, the actual test questions, the domain of objectives from which the test objectives are sampled, or the domain of items from which the test questions are sampled.

The public is concerned because they wanted accurate data from which to make inferences about the effectiveness of the schools/educators, but they keep hearing that the data are not any good because the teachers are cheating.

SOME PSYCHOMETRICIANS' VIEWS

In 1984 I attempted to assist those educators who wished to engage in some scholarly thinking by publishing an article (Mehrens, 1984) in which I stressed that one frequently wishes to make inferences to a reasonably broad domain. In such cases, a test can only sample the domain, and it is counter-productive to match the curriculum to the sample. This is true whether the sample is actual test items from the test bank domain or whether the sample is a set of objectives sampled from the domain of objectives. Following that article, Mehrens and Phillips published a series of analytical and empirical papers (Mehrens & Phillips, 1986, 1987; Phillips & Mehrens, 1987, 1988) related to the topic. In one of our papers, we took a previously developed matrix of 1260 cells which defined the domain of fourth grade mathematics and collapsed that matrix into a set of 180 cells. We matched an existing test with 53 questions to 53 of the 180 cells. Clearly the objectives and test questions in that test could only sample that domain. Another test might map

onto the 180 cell domain somewhat differently. However, both tests could be sampling the same domain. Certainly, if one taught to the specific 53 cells covered with the test we mapped, one could not make the inference from the test score to the domain.

Then, in 1989, Mehrens and Kaminski (1989) published a paper in which we presented a continuum of instructional strategies and located a point on that continuum where we felt one passed over the line from providing legitimate instruction to inappropriate instruction. Others have responded to that article. For example, while Mehrens and Kaminski thought it inappropriate (indeed unethical) to instruct on a parallel form of the same test, Bauernfeind (1989) has advocated such a practice. Cohen and Hyman wish there were a "national cheating conspiracy" (1991, p. 20). They "choose to confront NRTs as one more social demand that we must train children to cope with" (p. 23). Mehrens (in press) has responded that: "To intentionally set out to invalidate the inferences users wish to draw from tests seems to me to be unwise if not morally reprehensible."

Others have also entered into the discussion (see Hall & Kleine, 1990; and Nolen, Haladyna, & Haas, 1990). Popham (in press) has presented the issue using a slightly different approach and has, in my opinion, made a significant contribution by focusing, among other things, specifically on the instructional format.

MY CURRENT VIEWS

I should stress that my current views have been influenced by the views of others. While it is difficult to credit all who have influenced my thinking, I would particularly acknowledge Paul Sandifur, Jim Popham, and Tom Haladyna -- but not hold them responsible for my views.

1. The most general--and somewhat abstract--principle is that a teacher should not engage in any type of instruction that attenuates the ability to infer from the test score to the domain of knowledge/skill/or ability of interest.

2. It is appropriate to teach the content domain to which the user wishes to infer. This means that the domain must be defined. Reasonable people can and do disagree about how precise the definition must be. I believe that giving samples of detailed item specifications -- as was done in the TECAT (see Shepard, 1987) -- is inappropriate because the inference of interest is not simply whether students can master the domain when the questions are asked in a very specific format.

3. It is appropriate to teach test-taking skills. Both generic skills and those appropriate to a particular test item format may be taught. However, it is not necessary to spend an inordinate amount of time instructing students on how to take tests.

4. It is inappropriate to limit content instruction to a particular test item format. One should not teach science, mathematics, or any other subject by focusing student activities toward a particular test format.

5. It is inappropriate to teach only the objectives from the domain that happen to be sampled on the test. Teaching to only the sample tested when the inference is to a broader domain invariably makes the inference invalid. (It is because parallel tests are typically not randomly parallel, but test over the same sampled objectives, that I view it as inappropriate to have students practice on parallel tests. If the parallel tests both covered separate random samples from the domain, it would not disturb me to have them do that as long as those parallel "practice" forms were never used later as "live" tests.)

6. It follows from the above that it is inappropriate to use any commercial or locally developed instructional guide that "provides your students with the concentrated practice and review of the very skills necessary to score high on the latest edition of each test" (emphasis added, Random House, 1987, pp. 2-3).

7. It is definitely inappropriate to limit instruction to the actual test questions themselves.

8. If the test objectives comprise (i.e. do not sample from) the domain objectives, it is obviously appropriate to teach directly toward the test objectives. For example, we should teach students to recognize all 26 letters in the English alphabet, even if it turns out that all 26 letters are on some test. (However, it is seldom that we are interested in such a narrow domain that the test objectives comprise the domain objectives.)

9. If one notes that a test uses particular vocabulary in its instructions, it is appropriate to make sure the students understand that vocabulary. For example, the instructions might use words like minuend and subtrahend in a substraction section of an arithmetic test. One should instruct the students so that they can understand what they are to do. It would, of course, be inappropriate to teach the meaning of those two words at the exclusion of other words in the domain, if they were in a section designed to assess the students' competence in mathematical vocabulary.

10. All the above points apply to any kind of assessment. Some advocates of performance assessment seem not to understand the domain/sample issue. However, for any assessment where the inference is to go beyond the specific task, one can not teach toward only the specific task. There may be some physical skills where the specific skill itself is what one wishes to make inferences about. Much more frequently, especially in the subject

matters where one wishes to assess what the cognitive psychologists refer to as procedural knowledge or metacognitions, one would simply be making an incorrect inference if the specific "performance task" was the object of instruction.

GREY AREAS

While I presently feel fairly comfortable with the above guidelines, there are some gray areas which need discussion. Two of them are as follows:

1. No set of high stakes assessment procedures will cover all the domains that the schools are interested in the students mastering. Thus, one could concentrate instruction on the total domain that one wishes to infer to from the test score and still be delimiting instruction inappropriately. The problem is alleviated somewhat if high-stakes assessments are quite broad.

2. One should use assessment results to improve both instruction and the curriculum. One would hope the improvement would result in greater student competence in the domain of interest. However, when one observes a weakness in a class/student on a particular objective, it is natural and wise to want to assist that class/student on that particular objective and to instruct future classes/students so as to alleviate the weakness. This may corrupt somewhat the inference to the domain. There must be some point of trade-off between improvement on some specific objectives at the expense of being able to make accurate domain inferences.

TANGENTIAL ISSUES

Should We Have High Stakes Tests?

Probably, but we must recognize the negatives: (1) School efforts may be disproportionately expended on the domains (or worse yet the sampled

objectives or items) which are assessed--short-changing efforts on other useful educational goals. (2) There will be increased efforts to teach inappropriately to the assessments (both in terms of the sampled objectives/items as well as the format of the assessments), thus corrupting the meaning of the scores. (3) To the extent the efforts toward corruption succeed, neither the public nor the educators will make correct inferences from the data. (4) This lack of correct inferences will hinder public and educational efforts to improve education.

The Responsibilities of Those Who Mandate/Build Tests

The NCME task force on enhancing the credibility of school testing programs addresses this issue a bit. Drawing from a draft of that report and adding some of my own ideas, I suggest at least the following: Those who mandate and/or build tests must (1) define the domain which they believe is sampled by the test, (2) describe the test preparation activities (if any) used by the norming schools, and (3) describe and defend what they believe to be appropriate test preparation activities (those that would not impede accurate inferences to the domain of interest).

Is This Issue Relevant in Performance Assessment?

Unfortunately, as mentioned earlier, some seem to believe (incorrectly) that the issue is not relevant to performance assessment. However, if anything, the issue is even more important for a variety of reasons. The same problems exist as to whether the correct domain is being assessed and whether it is well defined. The sampling problems are greater in performance assessment and typically the domain will not be adequately sampled. Finally, if one teaches students how to respond to specific "higher order thinking

skills" or a "metacognition assessment" by teaching them to memorize the correct answers (performances) one cannot even make the correct inference to the sample.

REFERENCES

Bauernfeind, R.H. (1989). Article on test taking problematic. Educational Measurement: Issues and Practice, 8(4), 28.

Cohen, S.A. and Hyman, J.S. (1991). Can fantasies become facts? Educational Measurement: Issues and Practice, 10(1), 20-23.

Hall, J.L. & Kleine, P.F. (1990, April). Educator perceptions of achievement test use and abuse: A National Survey. Paper presented at the annual meeting of the National Council on Measurement in Education. Boston, MA.

Mehrens, W.A. (1984). National tests and local curriculum: Match or mismatch? Educational Measurement: Issues and Practice, 3(3) 9-15.

Mehrens, W.A. (in press). Facts about samples, fantasies about domains. Educational Measurement: Issues and Practice.

Mehrens, W.A. and Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless or fraudulent? Educational Measurement: Issues and Practice, 8(1) 14-22.

Mehrens, W.A. and Phillips, S.E. (1986). Detecting impacts of curricular differences in achievement test data. Journal of Educational Measurement, 23(3) 185-196.

Mehrens, W.A. and Phillips, S.E. (1987). Sensitivity of item difficulties to curricular validity. Journal of Educational Measurement, 24, 357-370.

Nolen, S.B., Haladyna, T.M., & Haas, N.S. (1990, April). A survey of actual and perceived uses, test preparation activities, and effects of standardized achievement tests. Paper presented at the joint annual meetings of the American Educational Research Association and the National Council for Measurement in Education. Boston, MA.

Phillips, S.E. and Mehrens, W.A. (1987). Curricular differences and unidimensionality of achievement test data: An exploratory analysis. Journal of Educational Measurement, 24(1), 1-16.

Phillips, S.E. and Mehrens, W.A. (1988). The effects of curricular differences on achievement test data at the item and objective levels. Applied Measurement in Education, 1(1) 33-51.

Popham, W.J. (in press). Appropriateness of teachers' test-preparation practices. Educational Measurement: Issues and Practice.

Shepard, L.A. (1987). A case study of the Texas Teacher Test: Technical report, Center for the Study of Evaluation, Graduate School of Education, University of California, Los Angeles.